**Assignment #2**

|  |  |
| --- | --- |
| **Name** | **ID** |
| Essam Alaa Nady | 20216064 |
| Ziyad Alfikie | 20216043 |
| Mariam Behairy | 20216094 |
| Mahitab Yasser | 20216082 |
| Ahmed El Tokhy | 20216014 |

**Problem 1 [Decision Trees]:**

**Introduction:**

The objective is to build a model that can predict the appropriate medication/Drug type (target variable) for patients suffering from a specific illness based on their Age, Sex, Blood Pressure, and Cholesterol levels.

**Data Processing:**

**1- Loading dataset:**

A screenshot of a computer

Description automatically generated

**2- Handling missing values:**

A screenshot of a computer

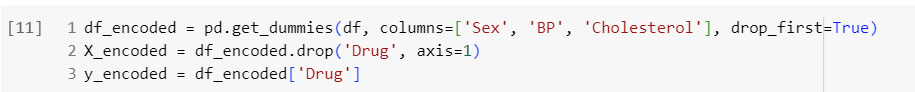
Description automatically generated

**3- Checking the type of each feature (categorical or numerical):**

A screenshot of a computer code

Description automatically generated

**4- Apply One-Hot Encoding for Categorical Features:**

****

The pd.get\_dummies() function is used to perform one-hot encoding, creating binary columns for each category.

**First Experiment:**

**A screenshot of a computer program

Description automatically generated**

This code performs the first experiment by splitting the data into training and testing sets five times with different random states. It trains a Decision Tree model on the training set and evaluates its accuracy on the testing set for each experiment. Finally, it prints out the accuracy of each experiment and the best-performing model of all experiments.

**Second Experiment:**

**1- Lists to store statistics:**

**A white background with green text

Description automatically generated**

Create lists to store the statistics for mean, max, and min accuracies, as well as mean, max, and min tree sizes. np.arange(0.3, 0.8, 0.1) generates an array of split ratios from 0.3 to 0.7 in increments of 0.1.

**2- Perform the experiment for each training set size:**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

For each split ratio, perform the experiment five times with different random seeds. Within each iteration: Split the data using the current split ratio, Initialize the Decision Tree Classifier, Train the model on the training set, Make predictions on the test set, Calculate accuracy and tree size for each experiment, and Store accuracy and tree size for each seed in separate lists.

Finally, Calculate the mean, max, and min statistics for accuracy and tree size.

**3- Create the report:**

**A computer code with black text

Description automatically generated**

**4- Print the report:**

**A screenshot of a computer

Description automatically generated**

**5- Accuracy vs Training Set Size Plot:**

**A screenshot of a computer screen

Description automatically generated**

**6- Tree Size vs Training Set Size Plot:**

**A screen shot of a computer

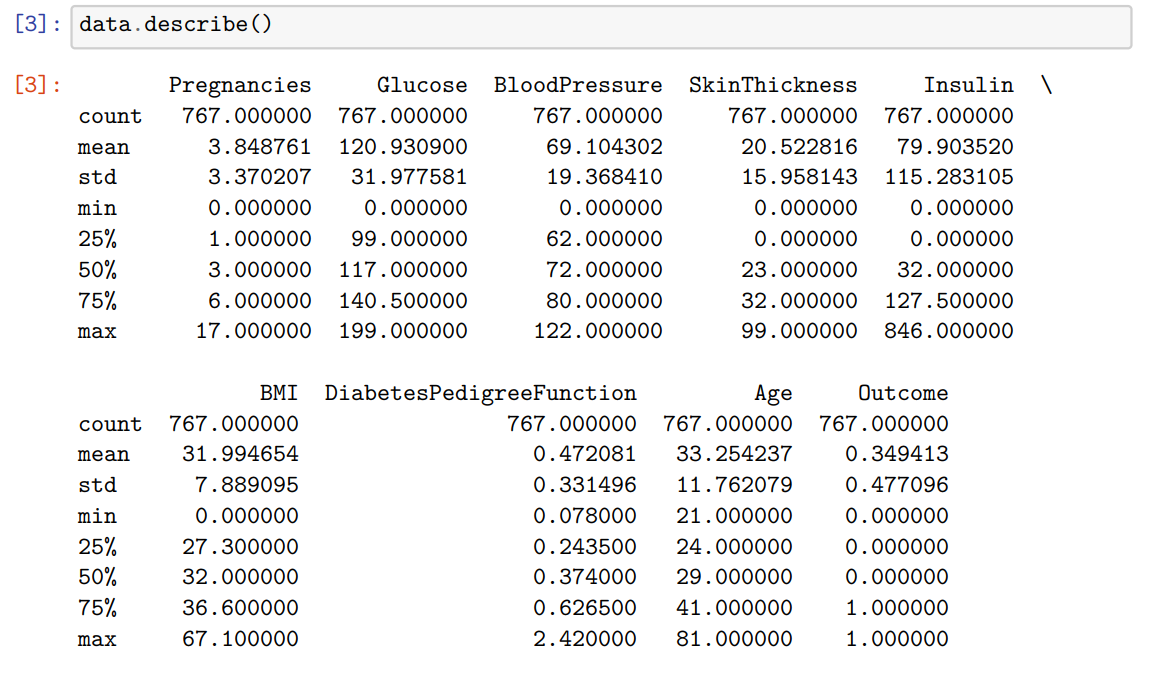
Description automatically generated**

**Problem 2 [KNN]:**

**Introduction:**

The objective is to build a model to implement your own simple KNN classifier.

**Exploring Data:**

****

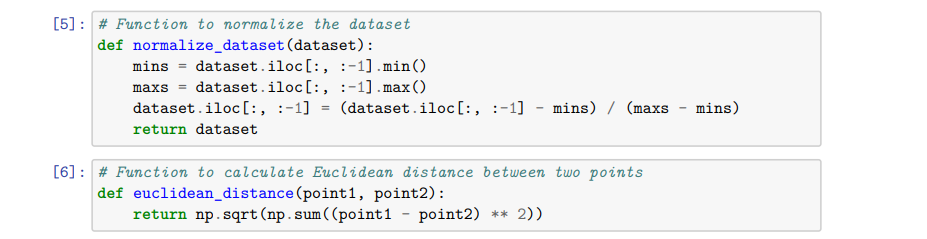
Data need to be scaled due to different ranges between features.

**A screenshot of a computer

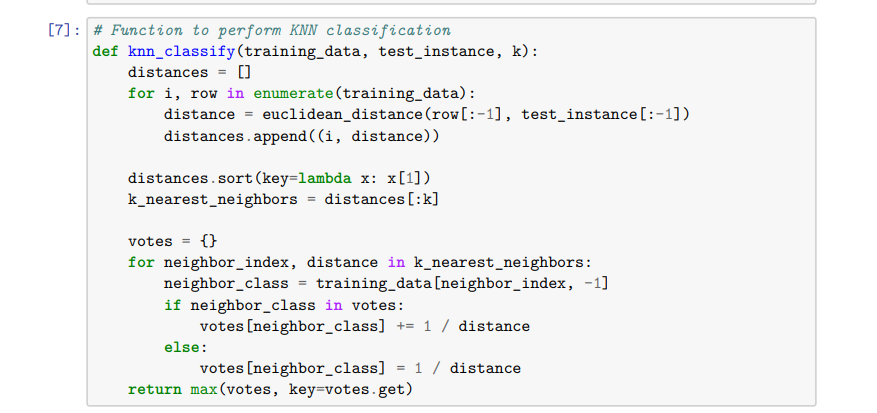
Description automatically generated**

No Null values in the data

**Implementing The KNN:**



Scaling data with min max scaler and calculating the Euclidean distances between instances.



Function to Perform the KNN by Calculating the Euclidean distances between each test point with all the trainings points then select the nearest K points then select the label with highest vote value.

A screenshot of a computer code

Description automatically generated

Splitting data into 30% testing an 70% training then scaling data using min max scaler

A screenshot of a computer code

Description automatically generatedA screenshot of a computer

Description automatically generatedperforming KNN on our data with different K values and print some accuracy information about each k value , from the output its obvious that we have better results when K = 5